Prologue to Perfectly Parsing Proxy Patterns

Jeremy M. Brown
Mercer University Libraries, brown_jm@mercer.edu

Gretchen M. Smith
Mercer University Library, smith_gm@mercer.edu

Jeremy M. Brown and Gretchen M. Smith, "Prologue to Perfectly Parsing Proxy Patterns" (2017). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284316683

Prologue to Perfectly Parsing Proxy Patterns

Jeremy M. Brown, Mercer University Libraries

Gretchen M. Smith , Mercer University Libraries

Abstract

As libraries spend an increasing percentage of precious collection funds on electronic resources, important questions arise to drive collection management decisions: What is being used? How much? and finally, Who is using our resources? Vendor-supplied statistics can help answer the first two questions, but we have encountered specific questions about our users at Mercer University.

To help answer this question, we turned to our proxy server logs and began a pilot study in the spring semester 2017. This presentation will explain the methodology we used in mining data from our proxy server logs in combination with our existing user database. It will describe the demographic information we were able to glean from this combination of information resources. We uncovered valuable insights to our database usage including: usage pattern over time, database popularity by program, database usage by enrollment status, usage by faculty/employee group, and usage by campus group.

Electronic resources present libraries with a number of attractive features. Among others, they promise easy access regardless of geography and tremendous searching advantages when compared to print resources. However, we have found that our faculty and sometimes administrators insist that their particular programs are not using these electronic resources, and perhaps exclusively rely upon print versions. We have many tools at our disposal to help compare usage of our electronic resources, but these fail to break down users by department, program, or other classification within the university.

Mercer University is a diverse institution in which electronic resources seem to offer some competitive advantages. We have major and satellite campus locations throughout Georgia. The university offers online and hybrid degree programs, and we are licensed to deliver distance learning in 42 states. Print resources are not a good fit for this geographically distributed teaching model. As one might expect, the University Library has been trending toward electronic resources and trimming our traditional print acquisitions and holdings in recent years.

Table 1. Extent of electronic vs. print resources.

E-journals	185,541
E-books	593,527
Total e-resources	947,628
Physical resources	370,572

The library has faced resistance against the use of electronic resources that can be broken down into two categories. In the first category, we see clientele lamenting the demise of the traditional library collection, and asking questions such as “Why are you making the library collection smaller?” This concern is often alleviated by demonstrating, as in Table 1, that we are actually expanding our holdings and access with digital materials. Our argument is further bolstered by the multiuser characteristic of the vast majority of these holdings. Whereas a single book or journal can only ever be used by a single user at a time, our electronic resources can virtually always be used simultaneously by multiple people. The second category is more complex, but it is basically a failure to recognize the quality or geographically neutral aspect of digital resources. This clientele is also concerned about library hours of operation as an impediment to access, or perhaps they require use of print resources in assignments. When we poll students, we have found that they come to the library as a place of study or for programming or other services more frequently than they visit to retrieve resources. Indeed, we only lend some 15,000 physical items per year (and many of these are reserve materials), but we see nearly a million annual accesses of our electronic resources. To address this criticism in the most effective way, we decided that we should construct a link between our users and the resources that they require.

https://doi.org/10.5703/1288284316683

Literature Review

To determine what other libraries have already done with proxy server logs to collect user data, we conducted a narrow literature review, limited to articles focused on electronic resource use that have been published since 2000.

The earliest articles, published before 2010, used data from large institutional surveys like the National Survey of Student Engagement. While these showed a correlation between library use and student success, they did not use library-specific data (Kuh & Gonyea, 2003; Laird & Kuh, 2005). This began to change following the publication of *The Value of Academic Libraries: A Comprehensive Research Review and Report* (Oakleaf, 2010). This report urged libraries to collect their own data to demonstrate the value of the library. Thereafter, most studies began using library-specific data.

Several articles proved useful in showing how to combine demographic data and usage data to show who was using library resources. Haddow and Joseph (2010) used enrollment data, demographic data, and library-use statistics to show a relationship between library use and student retention. Two other studies, Cox and Jantti (2012) and Stone and Ramsden (2013), used library-collected data in conjunction with demographic information to illustrate a correlation between library usage and academic performance.

The last set of articles dealt with how other libraries used proxy server logs to gather usage data about their users. Two studies conducted at the University of Minnesota Twin Cities provided invaluable information. Nackerud, Fransen, Peterson, and Mastel's (2013) article detailed the process for collecting data from the proxy server logs and looked at broad trends that the data showed. The subsequent article by Soria, Fransen, and Nackerud (2013) used the same data set but provided a more in-depth analysis and tied in demographic data to show a relationship between library usage and student GPA. Finally, Samson's study conducted at the University of Montana (2014) provided additional details about collecting data from the logs. It also addressed some of the issues we were facing, including how to gather information about on-campus users who do not need to authenticate to access electronic resources.

Methodology

When we considered the tools at our disposal, we realized that we had a rich set of data at our disposal. To construct user records for our automated library system, our campus Information Technology Department has created data extracts from our student management system and our human resources systems. We also have a rich set of highly detailed logs from our remote access system, EZproxy, and the developer documentation on OCLC's website.

Our user data is broken down into employee and student subgroups, which are very similar to each other. Each database uniquely identifies users by a synthetic ID, called the MUID. Employees are broken down by divisions and campuses. Sometimes the operational titles are useful, but frequently they are unique and thus unhelpful. Students are broken down into campus, grade level, programs, and program versions. The latter frequently will correspond to a particular degree within a given discipline. It is possible for a single user to appear in both databases, and it is possible for a student to be enrolled simultaneously in multiple programs. For these cases, we will only view a single record: the employee record first, and the student record with the latest expiration date (i.e., an enrollment with the final class date on December 15 will be used over one finishing on December 8). To merge these

Table 2. Mapping employee and student data onto a statistics table.

Employee Field	Student Field	Merged Field
Operational title	Program version (degree)	Version-title
Division	Program	Program-division
Campus	Campus	Campus
Static Text "employee"	Grade level	Level

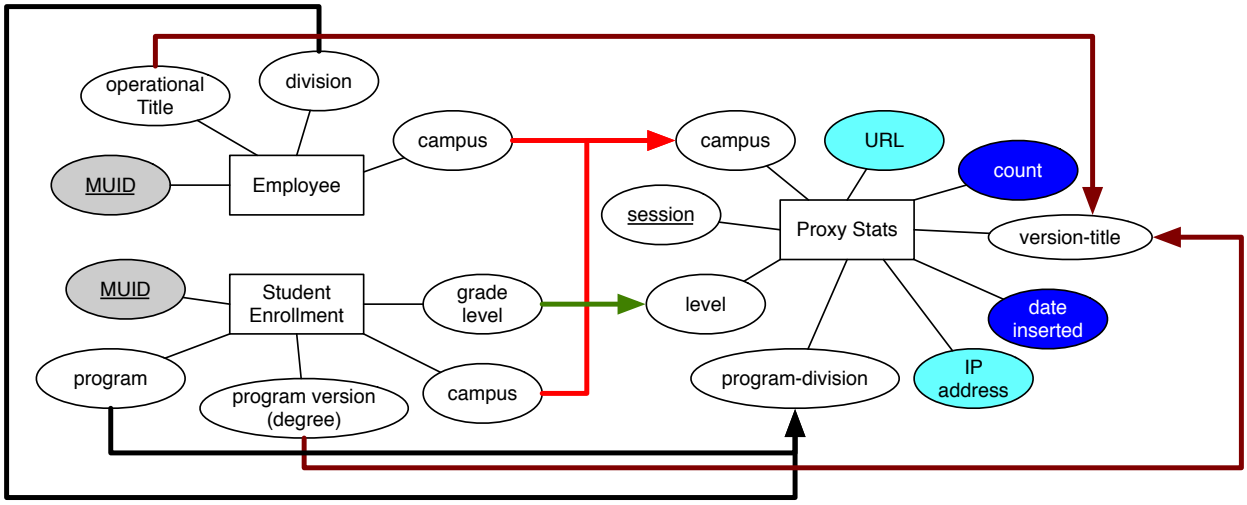

Figure 1. Data diagram depicting data available to us from our HR and Student Management systems and how we map it to the proxy statistics database.

databases, we decided to merge the fields depicted in Table 2. Figure 1 depicts these user databases and how they are mapped into the proxy statistics database. We discard the unique user identifier at this stage.

The main challenge in this paper is to connect this rich data about our user demographics with actual usage information from our EZproxy system. Fortunately, the EZproxy logs are easily understood and well documented (OCLC, 2015). We have deployed our instance of EZproxy utilizing one of the example log configurations:

> LogFormat %h %{ezproxy-session}i %u %t %r %s %b %{referer}

For this project, we utilized the elements detailed in Table 3, and we did not consider the remainder of the log line.

Table 3. Elements from the EZproxy log that we utilize and their meanings.

Variable	Meaning
%h	IP address of remote user
%{ezproxy-session}i	session ID
%u	username (our MUID)
%t	date/time request was made
%r	complete URL sent through

Our data ingestion process was worked into the log rotation process. Early each morning, we rotated the previous day's logs and started a new log file. We ran a small python script to parse each log entry and insert or update a row in the database. Users familiar with EZproxy will recognize that a typical EZproxy URL consists of a base URL (in our case http://proxy-s.mercer.edu) followed by a path, /login, and a query string that contains a particular resource's URL. For example: http://libraries.mercer.edu/login?url=http://library.artstor.org/. There are a number of redirects that happen as part of the log-in between EZproxy and our authentication page, but eventually we see a connect URL: http://proxy-s.mercer.edu/connect?session=s4hGslotZHbg&url=http://library.artstor.org/

This URL contains both a session identifier and the complete URL to the resource. We utilized this line to create a new row in our proxy statistics database. This gave us the URL, our unique identifier (the session), an IP address, and a user identifier. We then queried our user database for the demographics associated with the user identifier for the remaining elements of the database. To this, we automatically added a count of 1 and the current date. For each subsequent time this session identifier occurs in the EZproxy log, we incremented the count field by one.

This methodology was deemed sufficiently granular and effective, but it presented some limitations. The first is that in some cases we only saw the first database. For example, in an aggregated resource, such as ProQuest or EBSCOhost, a user could start

with one database and subsequently select one or more different products to search. We would never recognize that this had happened in our statistics. Another problem is that we did not isolate connect URLs within our aggregators, so all 66 of our EBSCOhost products look like the same resource. However, this can be easily remedied by studying the aggregator's URL documentation.

For the purposes of this pilot study, we analyzed data from March 7 through May 24, 2017. This period was roughly spring break through the end of the spring semester. We generated a number of views in our database to aggregate data and isolate particular variables. This data was then manipulated in Microsoft Excel using either simple tables or pivot tables and charts to visualize the results.

Results

We surveyed 38,491 data points using the four dimensions (campus, level, program-division, and version-title) from our proxy statistics database. We proceeded to create summary results tables, which give an overview of the entire survey period, and we also produced results by day. These daily results were visualized with pivot tables.

We could quickly see which programs utilized our electronic resources the most. Figure 2 gives the overview results, and it is evident that the College of Pharmacy makes up just over a third of our usage, followed by our "unknown" users at 23%. Figure 3 shows usage by day for the entire survey period. This daily stacked area chart shows a number of important usage milestones. We began our survey during spring break, which concluded on March 12, after which usage picked back up. We also saw a steep dip in activity in early May, which coincided with the end of the spring semester and the beginning of intersession. During the spring session, we also saw several peaks and valleys. The valleys generally coincide with weekends with a marked decrease in weekend activity. The chart also shows that the Pharmacy program sustains usage throughout holidays and into the summer, whereas our next largest group, our "unknowns," falls off dramatically during those periods.

We next turned to usage by grade level. As shown in Figure 4, the single largest grade level is "unknown," which is followed by our professional and graduate classifications. As one ascends in academic studies, it makes intuitive sense that one would conduct more research and utilize the library more heavily. This seems to be borne out by the facts in Figure 4 with a few exceptions; in particular, our master-level students use more electronic resources than our doctoral candidates. However, the master's students also outnumber the doctoral candidates by 10:1, which explains the seemingly higher usage. The same is true with "1st Year Graduate/Professional" at 9% of usage compared to the third-year students, who are at 7%: there are simply more students in their first year. One thing to note here is also that our undergraduates have a very meager presence in the data.

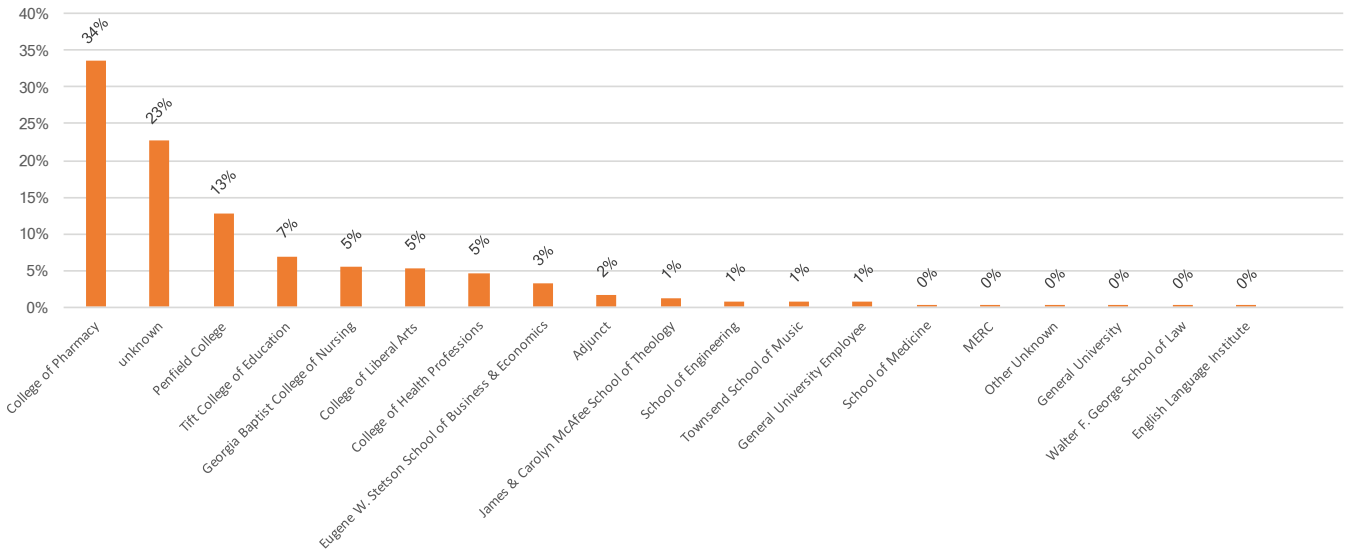

Figure 2. Usage results by program.

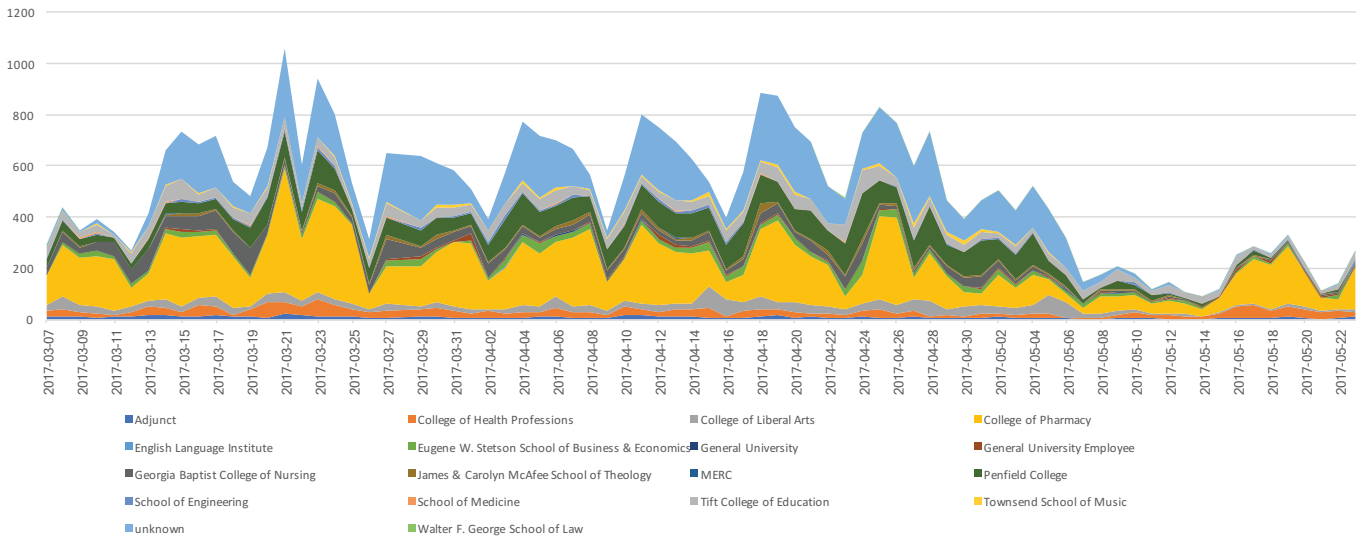

Figure 3. Usage results by program over time.

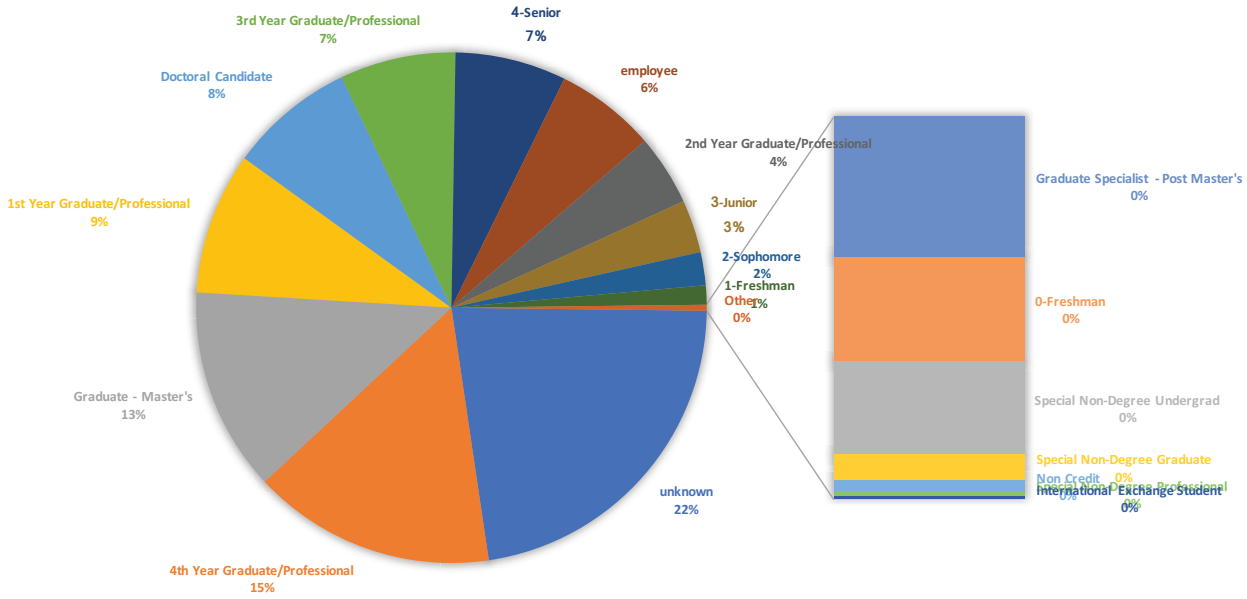

Figure 4. Usage results by level.

Utilization by campus is also an important figure, as Mercer has almost as many graduate/professional students on our Atlanta campus as we do on our traditional undergraduate campus in Macon. Our primarily nontraditional students are spread across our other locations as well as our "centers." We see in Figure 5 that Atlanta accounts for over half of our total electronic resource usage, perhaps because, as shown in Figure 2, the College of Pharmacy in Atlanta accounts for over a third of our usage by itself. Other heavy users are likewise located on the Atlanta campus: Tift College of Education, Georgia Baptist College of Nursing, and the College of Health Professions. It is important again to note how many of our users are unknown, some 22.47%, and how few users are from the Macon campus, 8.38%. Figure 6 demonstrates how steady our usage is. This trend continues until most classes are out of session, when our Atlanta campus completely dominates the stacked area chart. This tendency is evident in the data from spring break in early March and from after the end of the spring semester in May.

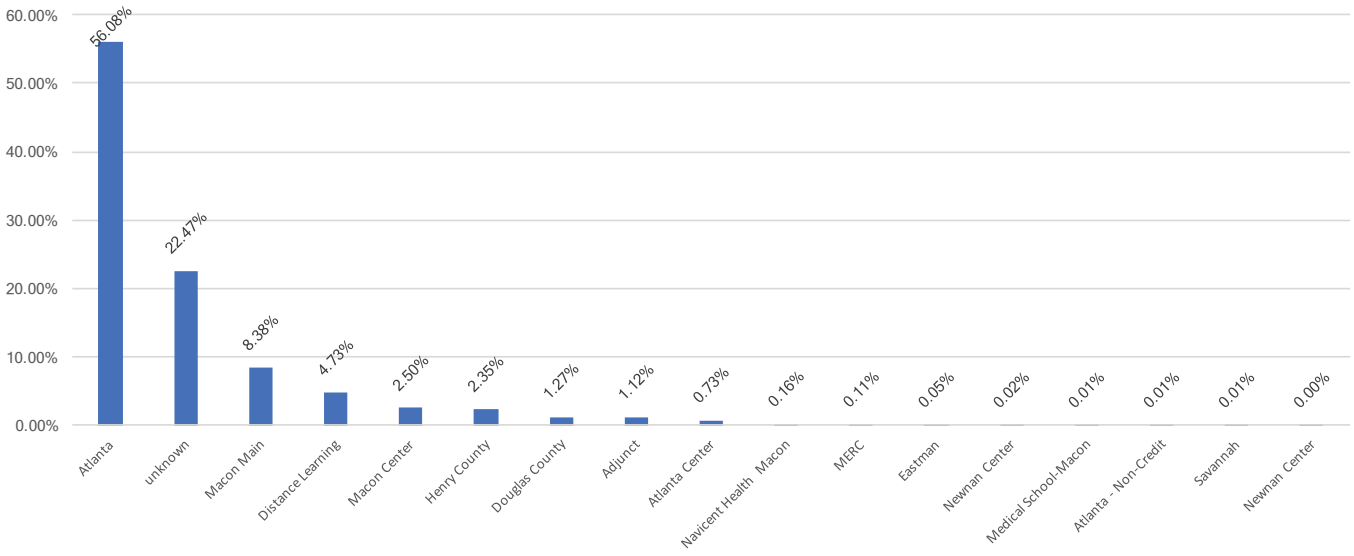

Figure 5. Usage results by campus.

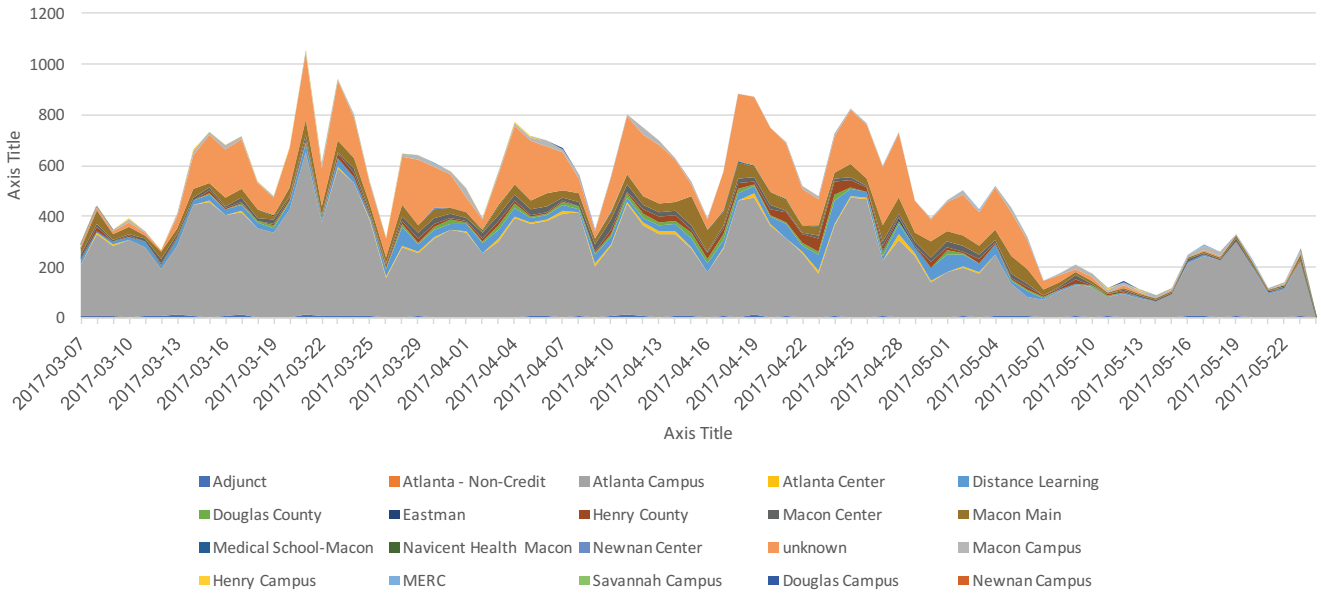

Figure 6. Usage results by campus over time.

Because our data has such a high percentage of unknown users, we decided to investigate this further. These users are "unknown" because we do not require on-campus users to log into the system. We know which database they use because we see the same "/connect" URL as mentioned previously, but the proxy is configured to simply direct them to the desired database URL. We do have a network map of our IP address ranges, and we can locate our users. This is illustrated in Figure 7, which shows that the Macon campus dominates this segment. The Macon undergraduate population accounts for nearly all of our "unknown" users, which stands to reason because there is a residency requirement. As a result, most students live on campus, so most students use the campus network to access resources. This trend also explains why these Macon undergraduates make up an unexpectedly small percentage of previous charts.

Our final investigation was to attempt to tie specific programs to particular database products. This endeavor is charted in Figure 8, where the outliers

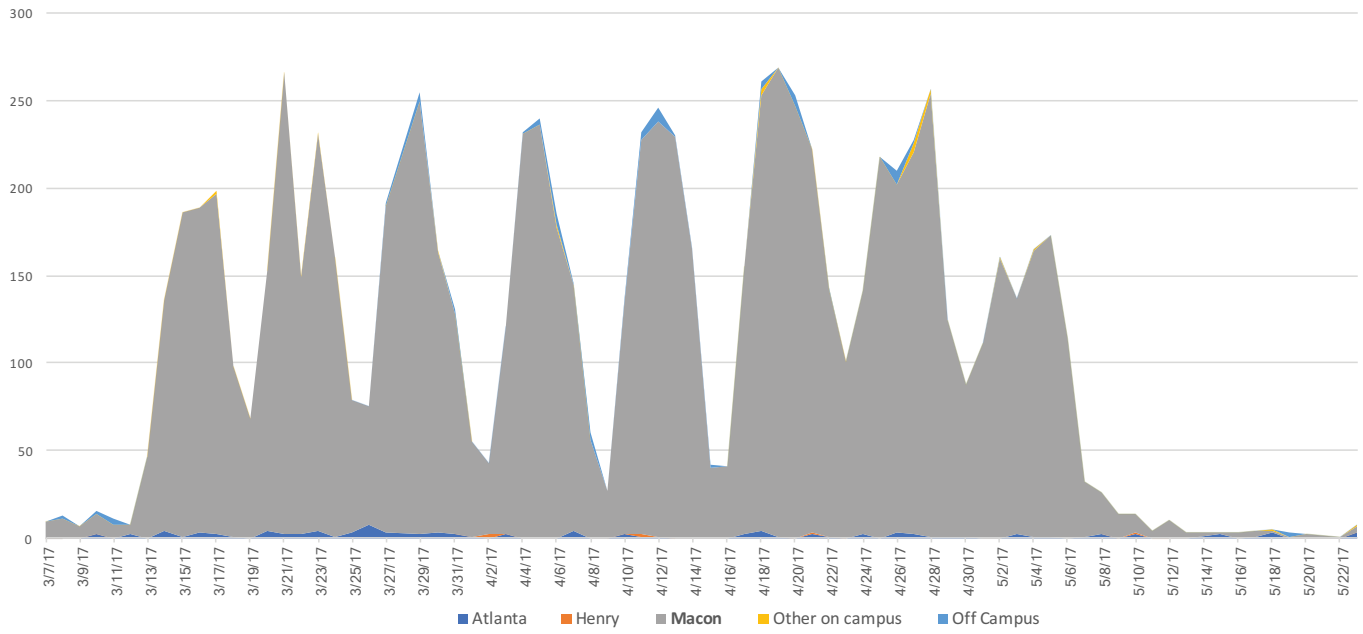

Figure 7. Usage results by unknown as determined by IP address.

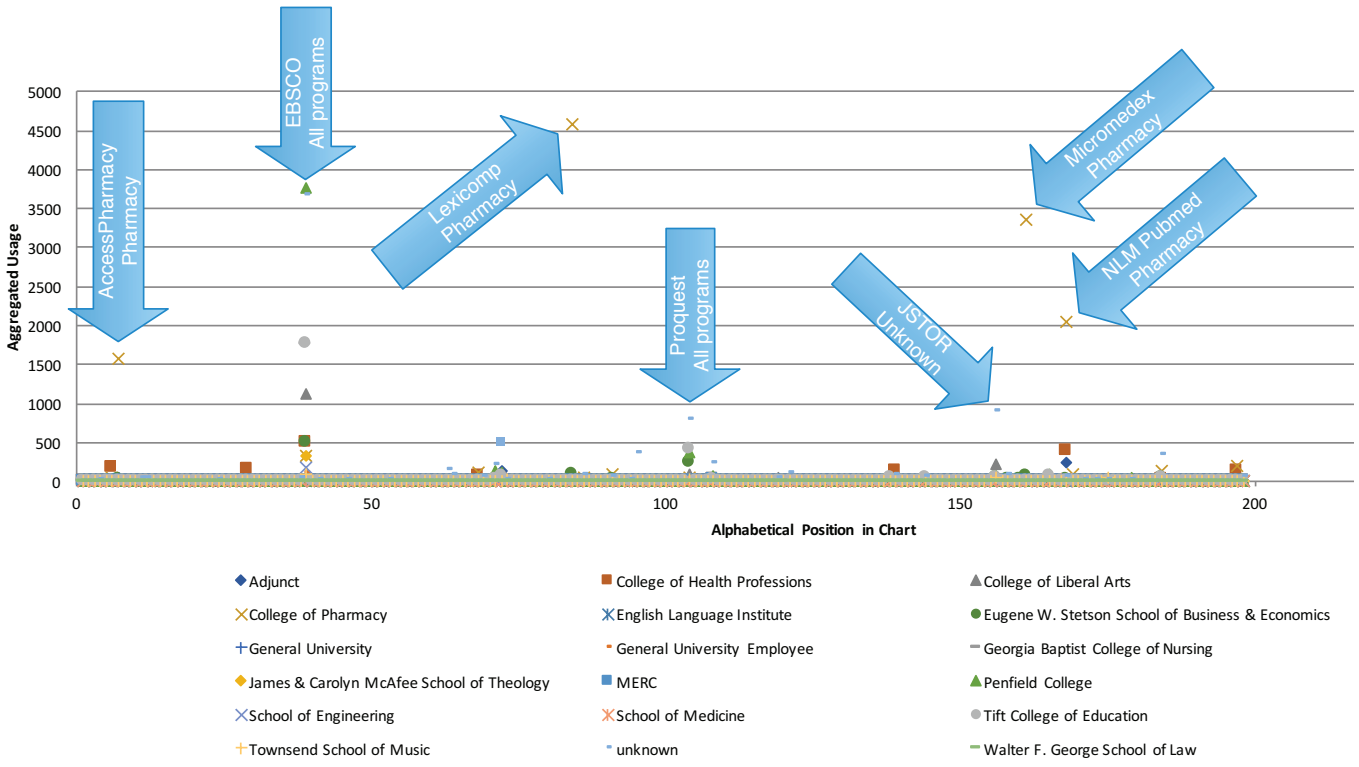

Figure 8. Usage results by database and program.

stand out from the background of relatively low utilization. The vast majority of outliers are either Pharmacy databases or databases that everyone uses, such as the aggregators EBSCOhost and ProQuest or JSTOR. The pivot table that generated this visualization is particularly insightful, as it allows us to investigate the actual numbers within the time span, showing how much, quantitatively, students and employees from any given program utilize any one product.

In conclusion, we believe that we have developed a tool to track database usage meaningfully by demographic group. For example, we could combine database usage for our School of Theology and say that they utilized our electronic resources some 300 times in this time period, but they only checked out some 250 physical resources. This sort of connection will allow us to guide future collection development efforts.

This project has some exciting implications for the future. We could combine this with grade data from our student management system and discover if student achievement correlates with library resource utilization. We could further begin tracking our other service data and add physical and virtual services into that analysis. We would also like to make a more granular database breakdown, so that our aggregators, like EBSCOhost and ProQuest, will not appear as single highly used databases, but rather as the individual databases they include.

References

Cox, B. L., & Jantti, M. (2012). Capturing business intelligence required for targeted marketing, demonstrating value, and driving process improvement. *Library and Information Science Research, 34*(4), 308–316. https://doi.org/10.1016/j.lisr.2012.06.002

Haddow, G., & Joseph, J. (2010). Loans, logins, and lasting the course: Academic library use and student retention. *Australian Academic & Research Libraries, 41*(4), 233–244.

Kuh, G. D., & Gonyea, R. M. (2003). The role of the academic library in promoting student engagement in learning. *College & Research Libraries, 64*(4), 256–282.

Laird, T. F. N., & Kuh, G. D. (2005). Student experiences with information technology and their relationship to other aspects of student engagement. *Research In Higher Education, 46*(2), 211–233. https://doi.org/10.1007/s11162-004-1600-y

Nackerud, S., Fransen, J., Peterson, K., & Mastel, K. (2013). Analyzing demographics: Assessing library use across the institution. *portal: Libraries and the Academy, 13*(2), 131–145.

Oakleaf, M. (2010). *The value of academic libraries: A comprehensive research review and report*. Chicago: Association of College and Research Libraries.

OCLC. (2015, March 2). LogFormat. Retrieved December 5, 2017, from http://www.oclc.org/support/services/ezproxy/documentation/cfg/logformat.en.html

Samson, S. (2014). Usage of e-resources: Virtual value of demographics. *Journal of Academic Librarianship, 40*(6), 620–625.

Soria, K. M., Fransen, J., & Nackerud, S. (2013). Library use and undergraduate student outcomes: New evidence for students' retention and academic success. *portal: Libraries and the Academy, (13)*2, 147–164.

Stone, G., & Ramsden, B. (2013). Library impact data project: Looking for the link between library usage and student attainment. *College & Research Libraries, 74*(6), 546–559.